BMC Bioinformatics

Research article

Evolutionary sequence analysis of complete eukaryote genomes
Jaime E Blair, Prachi Shah and S Blair Hedges*

Address: NASA Astrobiology Institute and Department of Biology, The Pennsylvania State University, 208 Mueller Laboratory, University Park, Pennsylvania 16802-5301, USA

Email: Jaime E Blair - jeb322@psu.edu; Prachi Shah - pss11@psu.edu; S Blair Hedges* - sbh1@psu.edu

* Corresponding author

Abstract

Background: Gene duplication and gene loss during the evolution of eukaryotes have hindered attempts to estimate phylogenies and divergence times of species. Although current methods that identify clusters of orthologous genes in complete genomes have helped to investigate gene function and gene content, they have not been optimized for evolutionary sequence analyses requiring strict orthology and complete gene matrices. Here we adopt a relatively simple and fast genome comparison approach designed to assemble orthologs for evolutionary analysis. Our approach identifies single-copy genes representing only species divergences (panorthologs) in order to minimize potential errors caused by gene duplication. We apply this approach to complete sets of proteins from published eukaryote genomes specifically for phylogeny and time estimation.

Results: Despite the conservative criterion used, 753 panorthologs (proteins) were identified for evolutionary analysis with four genomes, resulting in a single alignment of 287,000 amino acids. With this data set, we estimate that the divergence between deuterostomes and arthropods took place in the Precambrian, approximately 400 million years before the first appearance of animals in the fossil record. Additional analyses were performed with seven, 12, and 15 eukaryote genomes resulting in similar divergence time estimates and phylogenies.

Conclusion: Our results with available eukaryote genomes agree with previous results using conventional methods of sequence data assembly from genomes. They show that large sequence data sets can be generated relatively quickly and efficiently for evolutionary analyses of complete genomes.

Background

The use of complete genomes for phylogenetic analysis has greatly improved our understanding of prokaryote evolution [1-3]. However, until recently, relatively few complete genome sequences were available for such analyses in eukaryotes. As this improves, there will be a greater demand on methodology for evolutionary analysis of complete genomes. Previous whole-genome studies of eukaryotes have focused on gene and gene family pres-

ence-absence [4-7], lineage-specific gene loss [8,9], insertion-deletion markers and introns [6,10,11], and other non-sequence based information. While these approaches have their advantages, previous studies have not used complete genome sequences (nucleotides and/or amino acids) for reconstructing evolutionary relationships. At the same time, the complexity of eukaryote genomes, with numerous gene duplications and losses in different lineages, has created a challenge for sequence-based
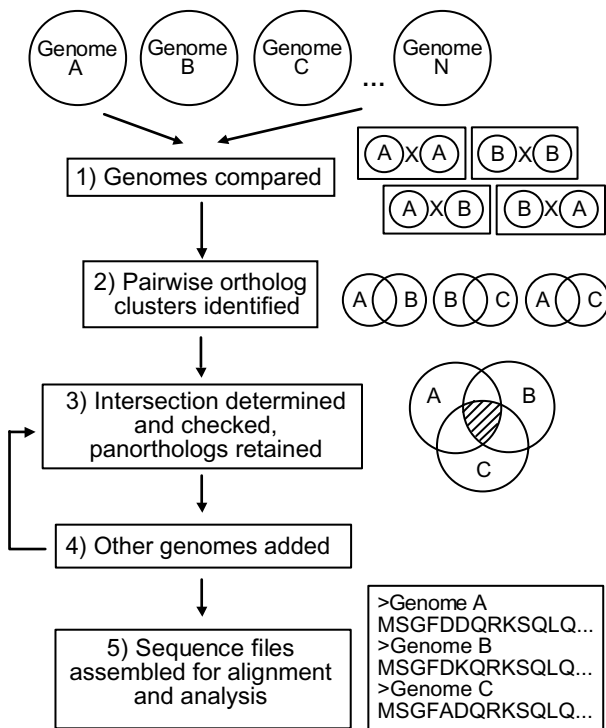
Figure 1
Flowchart of multigenome intersection approach (MIA). 1)
Complete genomes are reciprocally compared against themselves and all other genomes with BLAST. 2) Pairwise
ortholog clusters are identified using similarity scores and
imported into a local database. 3) The intersection between
genomes is determined by iteratively comparing sequence
identification tags and retaining those clusters showing panorthology. 4) Additional genomes are added and checked as
in the previous step. 5) Sequence data files are generated for
evolutionary analysis.

phylogeny estimation. Here, we outline a conservative
approach designed to utilize the wealth of evolutionary
information present in complete genome sequences by
identifying orthologs in multiple eukaryotes for the purpose of evolutionary analysis.

Methods for the identification of clusters of orthologs and
lineage-specific paralogs have proven useful for classifying
gene function and identifying cases where genes have
been differentially lost or duplicated in different lineages
[12-14]. However, such assemblages of data contain a
mixture of orthologs, paralogs, and missing data as a
result of gene loss, and are not generally suitable for large-scale phylogenetic sequence analysis of organismal evolution. Our approach for comparing multiple genome
sequences involves the identification of single-copy
orthologs across a number of genomes for evolutionary
analysis (Figure 1). We refer to such strict (1:1) orthologs

as *panorthologs*, in reference to their presumed "complete"
orthology, in contrast to *synorthologs*, which contain a mixture of species divergences and gene duplication events. In
other words, panorthologs are those genes (or clusters of
sequences) that contain only species divergences and do
not contain in-paralogs, out-paralogs, or co-orthologs
[15]. On the other hand, synorthologs are those genes (or
clusters of sequences) that contain species divergences
and any combination of paralogy (in-paralogs and out-paralogs). While the use of panorthologs is conservative
and reduces the number of usable genes or proteins, it
also lowers the probability that errors will be made in confusing a species divergence with a gene duplication event.
Because the ability to identify orthologs is diminished in
analyses of small to moderate numbers of species or
genomes, such a conservative method is appropriate in
those cases. This conservative approach has been used to
identify the number of shared, unduplicated proteins in
Homo sapiens, *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae*, where it was determined that such proteins perform primarily anabolic
rather than catabolic functions [16].

We compare our phylogenetic results and divergence time
estimates for an analysis of seventeen published eukaryote
genomes to a previous study that assembled nuclear protein sequence data in a more conventional manner from
public databases [17]. While the phylogenetic relationships between the organisms included in this study are
not controversial, with the exception of the position of
nematodes [18], this general approach will prove useful as
more genomes, including those with questionable phylogenetic affinity, are sequenced. In addition, this approach
facilitates the estimation of divergence times between
organisms with numerous molecular clock methods.

Results
The number of orthologous clusters per pairwise comparison and the percentage of those clusters showing panorthology are presented in Table 1. On average, pairwise
orthologous clusters contained approximately 60.3%
panorthologs; exceptions include comparisons between
fungi, including *Encephalitozoon* (average 89% panorthology), and all comparisons with *Arabidopsis* (average
34.6% panorthology). Comparisons within metazoans
averaged 54.7% panorthology, with *Mus* and *Rattus* showing the highest number of shared transcripts (16,413
orthologous clusters; 79.2% panorthology) as expected
due to their recent evolutionary divergence. Previous analyses showed approximately 12,400 panorthologs between
Mus and *Rattus* [19]. *Caenorhabditis elegans* and *C. briggsae*,
who diverged roughly 100 Ma [20], also shared a large
number of transcripts (12,416 orthologous clusters;
84.4% panorthology), which is similar to a previous estimate of 12,155 panorthologs [21]. The number of

Table 1: Number of orthologous clusters (upper-right) and percentage panorthologs (lower-left) per pairwise comparison[a].

|	Hsa	Mmu	Rno	Tru	Cin	Dme	Aga	Cel	Cbr	Sce	Ncr	Ago	Ecu	Ath	Cme	Gtn	Pfa
Hsa		14571	14201	9881	6100	5009	5081	3794	4114	1898	2197	1856	759	2892	1628	177	1198
Mmu	68.5		16413	9885	6115	4933	5112	4194	4182	1924	2240	1887	755	2894	1523	183	1241
Rno	73.2	79.2		9708	6016	5001	5023	4127	4079	1871	2207	1825	754	2836	1630	177	1235
Tru	61.6	61	67.5		4872	4970	4974	4109	4090	1464	2224	1808	725	2814	1591	177	1260
Cin	49.3	50.3	55.6	58.2		4520	4554	2980	3848	1823	2090	1740	700	2669	1494	170	1199
Dme	36.7	37.5	41.2	43.8	60.6		7072	3904	3822	1753	1967	1738	713	2476	1460	178	1141
Aga	41.3	41.7	47.4	51.5	72.4	61.3		3973	3926	1833	2107	1793	717	2641	1597	177	1214
Cel	34.2	37.4	42.2	45.7	65.1	49.9	58.8		12416	1549	1702	1593	697	2235	1368	171	990
Cbr	42	42.9	47.5	54.5	79.7	56.7	68.9	84.4		1611	1836	1561	691	2205	1348	168	1083
Sce	42.9	43.7	48.8	49.5	72.4	54.8	63.9	60.3	69		2604	4036	683	1867	1321	177	958
Ncr	48.3	47.3	52.8	59.2	83.1	59.2	71.9	67.9	77.9	84.3		2560	648	2182	1388	168	880
Ago	48.6	47.4	52.9	59.3	82.9	59.9	72.5	69.5	78.3	93.4	96.4		686	1818	1282	172	893
Ecu	42.6	39.5	43.5	55.7	77	55.4	66.8	68.4	76.4	77	91.8	91.3		741	567	143	525
Ath	29	28.5	32.2	32.4	40.9	33.3	37	32.9	36.4	36.6	39.3	38.6	33.1		2126	199	1295
Cme	48.2	45.4	52.1	59.6	83.2	57.6	71.3	68.5	77.7	78.7	90.4	92	90.8	40.9		198	849
Gtn	43.5	30.6	33.3	64.9	84.7	54.5	66.7	71.9	77.4	62.7	92.9	88.4	97.2	25.1	91.9		190
Pfa	46.2	42.1	46.9	58.7	80.9	57.9	71.6	71.4	78	79.6	91.6	91.3	91.8	36.6	91.2	94.7	

[a]Three-letter abbreviations listed in Methods.
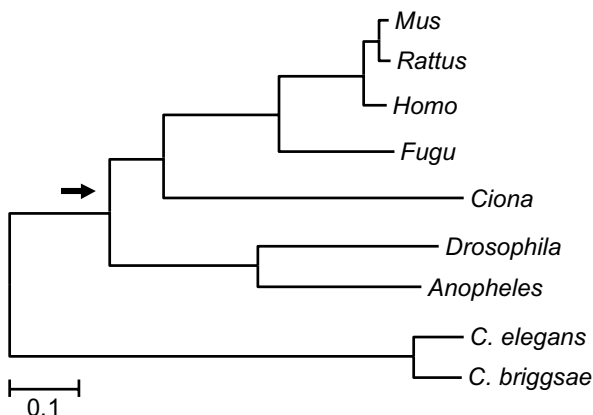
Figure 2
Neighbor-joining tree of nine metazoan genomes, 285 panorthologs (97,581 amino acid positions, alpha = 1.28). All nodes are supported significantly (>95%) in bootstrap analyses of neighbor-joining and maximum likelihood. The arrow indicates an alternative root [6, 18].

orthologous clusters between *Drosophila* and *Anopheles* found here (7072, 61.3% panorthology) is also similar to a previous estimate of approximately 6130 panorthologs [22]. Pairwise comparisons with the smallest genome, the *Guillardia* nucleomorph, averaged ~176 orthologous clusters, but the percentage of panorthologs varied greatly,

from a low of 25.1% with *Arabidopsis* to as high as 97.2% with *Encephalitozoon*.

The intersection of nine metazoan genomes resulted in a large number of shared genes. Among the nine genomes, 285 panorthologs were found, totaling 97,581 amino acids. The neighbor-joining tree of that concatenation is shown in Figure 2; all nodes in this tree received 100% bootstrap support. The intersection of all seventeen eukaryote genomes included in this study resulted in three shared genes (t-complex protein delta subunit, proteasome beta type-1 subunit, and Nip7p biogenesis factor) and orthology was confirmed manually. The reconstructed trees for the three genes showed long-branch attraction errors associated with the intracellular parasite *Encephalitozoon* and the *Guillardia* nucleomorph (data not shown). This was expected because both have highly reduced genomes and high rates of substitution across many genes as a result of their current or ancestral parasitic and symbiotic lifestyles [23,24]. For this reason, the intersection of the remaining fifteen genomes was determined, resulting in ten panorthologs. The intersection of genomes from twelve multicellular eukaryotes resulted in 63 panorthologs. The functional classifications of the panorthologs found here are similar to those identified in previous studies [14], with the most frequently represented functions being transcription, translation, replication and repair, and RNA processing. The phylogenetic trees reconstructed from the concatenated datasets both showed the expected relationships (Figure 3a and 3b) [17]. All nodes in these trees received very high bootstrap support, with only one node showing less than 95% boot-
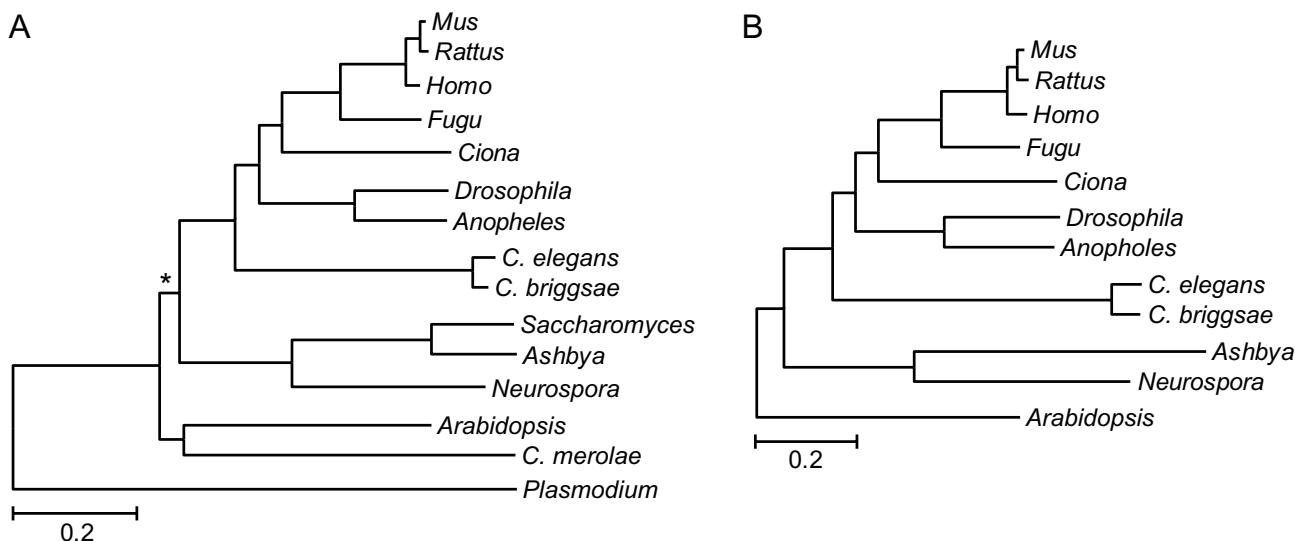
Figure 3
Neighbor-joining trees of complete eukaryotic genome sequence analyses. (A) The intersection of fifteen eukaryotic genomes, 10 panorthologs (5094 amino acid positions, alpha = 1.01). (B) The intersection of genomes from twelve multicellular eukaryotes, 63 panorthologs (23,571 amino acid positions, alpha = 1.15). All nodes are supported significantly (>95%) in bootstrap analyses of neighbor-joining and maximum likelihood, with the exception of node indicated by an asterisk (94% with maximum likelihood) in (A).

Table 2: Bayesian divergence time estimates (± one standard deviation) for eukaryote genome datasets.

Divergence	15 Genomes (5094 aa)[a]	12 Genomes (23,571 aa)	7 Genomes (132,190 aa)	4 Genomes (287,000 aa)
Mus – Rattus	37 ± 5	50 ± 8	40 ± 9	n/a[c]
Primate – Rodent	93 ± 10	117 ± 15	120 ± 20	n/a
Tetrapod – Fish[b]	459 ± 20	460 ± 20	460 ± 20	458 ± 20
Vertebrate – *Ciona*	771 ± 47	756 ± 58	n/a	n/a
Drosophila – Anopheles	445 ± 38	500 ± 48	477 ± 51	n/a
Chordate – Arthropod	949 ± 66	912 ± 89	955 ± 92	1109 ± 103
C. elegans – C. briggsae	89 ± 13	114 ± 22	n/a	n/a
Coelomata – Nematoda	1166 ± 89	1074 ± 116	n/a	n/a
Ashbya – Saccharomyces	311 ± 39	n/a	n/a	n/a
Saccharomycetes – *Neurospora*	900 ± 80	851 ± 96	n/a	n/a
Animal – Fungi	1493 ± 125	1303 ± 155	n/a	n/a
Arabidopsis – C. merolae	1414 ± 121	n/a	n/a	n/a
Animal/Fungi – Plant	1671 ± 145	n/a	n/a	n/a

[a]Number of amino acids (aa). [b]Tetrapod – actinopterygian fish divergence constrained between 425 and 495 Ma [78]. [c]Divergence not available (n/a) for timing.

strap support (animals + fungi in Figure 3a). The long branch observed in *Plasmodium* (Figure 3a) may be the result of both the long evolutionary separation from the other eukaryotes included in this study, and the high (A-T) composition of the genome [25] leading to biased amino acid compositions among proteins [26].

Phylogenetic trees were also reconstructed for each panortholog to test for congruence with well-supported phylogenies from the concatenated data (see Additional file 1). We found that in most cases, the consensus values calculated from individual trees agree with the high bootstrap support of the concatenated analysis. Two
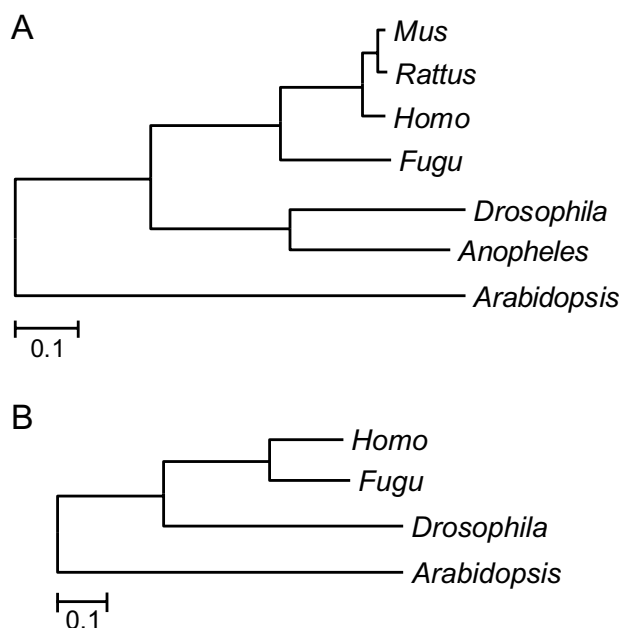
Figure 4
Neighbor-joining trees of genomes used to address deuterostome-arthropod divergence time. (A) The intersection of seven eukaryotic genomes, 380 panorthologs (132,190 amino acid positions, alpha = 1.38). (B) The intersection of four eukaryotic genomes, 753 panorthologs (287,000 amino acid positions, alpha = 1.46). All nodes are supported significantly (>95%) in bootstrap analyses of neighbor-joining and maximum likelihood.

exceptions were the low consensus values for the accepted close relationship between animals and fungi and the contested position of nematodes. Both taxa showed slightly longer branch lengths, and long-branch attraction artifacts [27] may be affecting the individual datasets, causing low consensus values. Also, recent empirical [28] and simulation [29] studies suggest that results from multigene analyses are more accurate when a tree is derived from a concatenated dataset of individual genes rather than a consensus of trees from multiple analyses.

Divergence times were estimated for both the 15-genome and 12-genome datasets (Table 2). Results were consistent with previous studies [17,30-33], showing an early divergence between plants, animals, and fungi (animals/fungi vs. plants ~1670 Ma, animals vs. fungi ~1400 Ma), and a Precambrian origin for animals (~900–1100 Ma). To specifically address the deuterostome-arthropod divergence within animals, two additional datasets were assembled to maximize the number of proteins analyzed: the intersection of seven genomes (*Homo*, *Mus*, *Rattus*,

Takifugu, *Drosophila*, and *Anopheles*; *Arabidopsis* as outgroup) and the intersection of four genomes (*Homo*, *Takifugu*, and *Drosophila*; *Arabidopsis* as outgroup). The seven genome intersection contained 380 panorthologs (132,190 amino acids; Figure 4a), and yielded a vertebrate-arthropod divergence time of 955 Ma. The four genome intersection contained 753 panorthologs (287,000 amino acids; Figure 4b) and yielded a vertebrate-arthropod divergence time of ~1100 Ma. Although this last estimate was derived from more than five times the data previously used, the divergence time is remarkably consistent with previous large-scale studies [17], and suggests that bilaterian animals originated hundreds of millions of years before the first fossil evidence of their existence in the Cambrian (~520 Ma). With the exception of the maximum fossil-based time estimate used in the tetrapod-actinopterygian fish calibration, the other fossil constraints used here are minimums, and therefore the resulting time estimates are minimums [34]. The agreement between our results and those of previous studies using different methods of data assembly suggests that our genome intersection approach is correctly assembling orthologs. Younger time estimates of the vertebrate-arthropod divergence have been obtained in some studies [35-37]. However, those results are problematic because they also contain estimates which are inconsistent (too young) with undisputed aspects of the fossil record, such as the oldest red algae (1200 Ma), green algae (1000 Ma), and stramenopiles (1000 Ma) [38-40]. Possible reasons for their inconsistency are discussed elsewhere [41].

Discussion
The complete genome sequence of an organism is essentially the maximum amount of discrete, genetically encoded information available for evolutionary analyses. However, orthology determination has been a major obstacle in the analysis of complete genomes, especially in eukaryotes where considerable gene duplication and loss has created additional complexity. Our approach for evolutionary analysis of complete eukaryote genome sequences is both simple and fast compared with the conventional method of gene-by-gene orthology determination using similarity searches in the public databases. The results of this approach applied to a subset of the available eukaryote genomes show agreement with previous results using conventional (non-genomic) approaches. In addition, the relatively high consensus values for most nodes indicate general agreement in tree topology among individual panorthologs.

The relatively low number of common genes in our intersections of 12–17 genomes is a combination of using panorthologs and including distantly related species. Genes are more likely to duplicate over long periods of evolutionary time, as in the time elapsed since plants separated

from animals (~1600 Ma) [17,32]. Therefore, a better approach with such distantly related species (e.g., all eukaryotes), and those groups with high levels of gene duplication and gene loss (e.g., nematodes), would be to relax the orthology criterion and include synorthologs. In that case, a representative or consensus sequence may be chosen from among in-paralogs. On the other hand, the implementation used here (panorthologs) should yield many genes in analyses of genomes from closely related species (e.g., within mammals), even if large numbers of species are used.

The use of sequence data for comparative genomics and phylogenetics has several advantages over the use of datasets based on the presence and absence or position of genes, introns and insertions. Sequence data can provide a larger number of characters for analysis, yielding hundreds of thousands of amino acid sites and more than a million nucleotide sites in some cases. Also, statistical models of sequence change are better known than those for non-sequence based data. Finally, the assembly of sequence data from complete genomes of multiple organisms not only facilitates phylogenetic and divergence time analyses, but a diversity of other comparative evolutionary analyses requiring sequence information [42,43].

Conclusion

Unlike previous studies of complete eukaryote genomes, here we have used a fast, conservative approach to assemble orthologous clusters of proteins for phylogenetic analysis and divergence time estimation. Our results are similar to previous studies that used conventional (slower) gene-by-gene data mining. We find that complete genome sequences support the close evolutionary relationship between animals and fungi, and also that molecular divergences between animals occurred approximately 400 million years before the Cambrian Explosion of fossils. Our approach will be further tested as more eukaryote genomes are sequenced.

Methods

Multigenome intersection approach for evolutionary analysis (MIA): Reciprocal BLAST [44] searches of genomes versus themselves and versus all other genomes included in the analysis were used to generate lists of pairwise similarity scores for each transcript. These scores were then used to generate orthologous clusters of proteins by first determining the "primary" ortholog pair through reciprocal best hits, then adding lineage-specific paralogs (in-paralogs) as implemented in INPARANOID [45]. The settings used here (sequence overlap cut-off 50%, group overlap cut-off 50%, in-paralog confidence cut-off 5%) were considered optimal in the sense that all closely related lineage-specific paralogs (and alternative transcripts) will be placed in the same pairwise cluster, mini-

mizing the probability that the same gene will be represented in more than one cluster. In-paralogs are added to clusters if they are more similar to one member of the primary ortholog pair than the two primary orthologs are to each other [45].

Only pairwise ortholog clusters can be obtained using INPARANOID. For phylogenetic analysis, ortholog sets for a larger number of genomes (at least four) must be constructed. Therefore, we combined the pairwise ortholog clusters from groups of species using a relational database. The intersection between ortholog clusters was determined by iteratively comparing sequence identification tags and reducing the intersected clusters at each round to exclude clusters that represent relationships other than panorthology. For example, consider genomes A, B and C. First, ortholog clusters are determined for each pairwise genome comparison, which results in clusters: A-B, B-C and A-C. Intersection of sets A-B and B-C is obtained by searching for common sequences of genome B in the two sets and merging the two sets accordingly into an ortholog cluster set of A-B-C. This combined set is reduced to include only clusters with panorthology relationships, i.e. only clusters with one sequence for each of the genomes A, B and C are retained. Further, the combined cluster A-B-C is compared with the pairwise cluster A-C based on common sequence tags for genome C. Any cluster from the combined A-B-C set that does not agree with the sequence grouping of genomes A and C is removed. This last step serves as an important check for orthology in each iteration of the intersection procedure, and is similar to the construction of three-member COGs (clusters of orthologous groups) [46]. The steps described above were performed iteratively in order to add more species to the ortholog clusters. Any number of genomes can be intersected (tested up to seventeen here), and an outgroup can be treated as part of the intersection or added separately by using a pairwise comparison to one of the in-group taxa. All programming was written in Perl.

Analysis of Eukaryote Genomes: Complete protein transcripts were obtained for the following eukaryote genomes [three letter abbreviation]: *Homo sapiens* [Hsa] (version 34b.2) [16,47,48]; *Mus musculus* [Mmu] (version 32.2) [48,49]; *Rattus norvegicus* [Rno] (version 3b.1) [19,48]; *Takifugu rubripes* [Tru] (version 3.0) [50,51]; *Ciona intestinalis* [Cin] (version 1.0) [51,52]; *Drosophila melanogaster* [Dme] (version 3.1) [53,54]; *Anopheles gambiae* [Aga] (version 2a.2) [48,55]; *Caenorhabditis elegans* [Cel] (version 120) [56,57]; *Caenorhabditis briggsae* [Cbr] (version 25) [21,57]; *Saccharomyces cerevisiae* [Sce] [58,59]; *Neurospora crassa* [Ncr] (version 3) [60,61]; *Ashbya gossypii* [Ago] (version 1.0) [62,63]; *Encephalitozoon cuniculi* [Ecu] [24,64]; *Arabidopsis thaliana* [Ath] (version 5.0) [65,66]; *Cyanidoschyzon merolae* [Cme] [67,68]; *Guil-*

lardia theta nucleomorph [Gtn] [23,64]; and *Plasmodium falciparum* 3D7 [Pfa] (version 4.1) [25,69]. Some genome transcripts were given unique sequence identifiers to avoid redundancy when sequence tags are truncated. The intersection of these seventeen genomes was determined as described above.

Each panortholog was aligned [70] and individual datasets were concatenated. Both individual panorthologs and concatenations were analyzed using maximum likelihood [71] to determine alpha parameters for the gamma rate-heterogeneity correction. Phylogenetic trees of concatenated datasets were reconstructed using neighbor-joining (Poisson + gamma correction model) with bootstrapping (2000 replicates) [72] and maximum likelihood (JTT + gamma correction model) with 1000 puzzling steps [73]. Phylogenetic trees of individual datasets were reconstructed using maximum likelihood (Poisson + gamma correction model) [71] and a consensus tree was derived [74]. Consensus values (i.e. the proportion of trees recovering a specific node) were calculated for each dataset.

Divergence times were estimated for concatenated datasets using Bayesian inference (JTT model) [75] as described previously [17]. The following fossil dates were used as minimum constraints: *Mus-Rattus* 12 Ma [76], primate-rodent 65 Ma [77], tetrapod-actinopterygian fish 425 Ma (lower bound) and 495 Ma (upper bound) [78], vertebrate-urochordate 520 Ma [79], *Drosophila-Anopheles* 250 Ma [80], chordate-arthropod 545 Ma [77], green algae/plants-red algae 1200 Ma [40].

Authors' contributions
JEB and PS designed and developed the methodology. PS programmed the genome intersection software. JEB carried out the evolutionary analyses and drafted the manuscript. SBH coordinated the research and assisted with drafting the manuscript.

Additional material

Additional File 1
Consensus trees of individual gene trees. Consensus trees of individual gene (panortholog) trees, showing percentage of individual gene trees supporting each node.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-6-53-S1.pdf]

Acknowledgements

References
1. Wolf YI, Rogozin IB, Grishin NV, Tatusov RL, Koonin EV: **Genome trees constructed using five different approaches suggest new major bacterial clades.** *BMC Evolutionary Biology* 2001, **1**:8.
2. Raymond J, Zhaxybayeva O, Gogarten JP, Gerdes SY, Blankenship RE: **Whole-genome analysis of photosynthetic prokaryotes.** *Science* 2002, **298**:1616-1620.
3. Brochier C, Forterre P, Gribaldo S: **Archaeal phylogeny based on proteins of the transcription and translation machineries: tackling the *Methanopyrus kandleri* paradox.** *Genome Biology* 2004, **5**:R17.
4. Korbel JO, Snel B, Huynen MA, Bork P: **SHOT: a web server for the construction of genome phylogenies.** *Trends in Genetics* 2002, **18**:158-162.
5. House CH, Runnegar B, Fitz-Gibbon ST: **Geobiological analysis using whole genome-based tree building applied to the Bacteria, Archaea, and Eukarya.** *Geobiology* 2003, **1**:15-26.
6. Wolf YI, Rogozin IB, Koonin EV: **Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis.** *Genome Research* 2004, **14**:29-36.
7. Copley RR, Aloy P, Russell RB, Telford MJ: **Systematic searches for molecular synapomorphies in model metazoan genomes give some support for Ecdysozoa after accounting for the idiosyncrasies of *Caenorhabditis elegans*.** *Evolution and Development* 2004, **6**:164-169.
8. Krylov DM, Wolf YI, Rogozin IB, Koonin EV: **Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution.** *Genome Research* 2003, **13**:2229-2235.
9. Hughes AL, Friedman R: **Differential loss of ancestral gene families as a source of genomic divergence in animals.** *Proceedings of the Royal Society of London B: Biological Sciences* 2004, **271**:S107-109.
10. Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV: **Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution.** *Current Biology* 2003, **13**:1512-1517.
11. Coghlan A, Wolfe KH: **Origins of recently gained introns in *Caenorhabditis*.** *PNAS* 2004, **101**:11362-11367.
12. Lee Y, Sultana R, Pertea G, Cho J, Karamycheva S, Tsai J, Parvizi B, Cheung F, Antonescu V, White J, Holt I, Liang F, Quackenbush J: **Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA).** *Genome Research* 2002, **12**:493-502.
13. Li L, Stoeckert CJJ, Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Research* 2003, **13**:2178-2189.
14. Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Rogozin IB, Smirnov S, Sorokin AV, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA: **A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes.** *Genome Biology* 2004, **5**:R7.
15. Sonnhammer ELL, Koonin EV: **Orthology, paralogy and proposed classification for paralog subtypes.** *Trends in Genetics* 2002, **18**:619-620.
16. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T,

Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, Szustakowki J, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
17. Hedges SB, Blair JE, Venturi ML, Shoe JL: **A molecular timescale of eukaryote evolution and the rise of complex multicellular life.** *BMC Evolutionary Biology* 2004, **4:**2.
18. Blair JE, Ikeo K, Gojobori T, Hedges SB: **The evolutionary position of nematodes.** *BMC Evolutionary Biology* 2002, **2:**7.
19. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE, Okwuonu G, Hines S, Lewis L, DeRamo C, Delgado O, Dugan-Rocha S, Miner G, Morgan M, Hawes A, Gill R, Celera, Holt RA, Adams MD, Amanatides PG, Baden-Tillson H, Barnstead M, Chin S, Evans CA, Ferriera S, Fosler C, Glodek A, Gu Z, Jennings D, Kraft CL, Nguyen T, Pfannkoch CM, Sitter C, Sutton GG, Venter JC, Woodage T, Smith D, Lee HM, Gustafson E, Cahill P, Kana A, Doucette-Stamm L, Weinstock K, Fechtel K, Weiss RB, Dunn DM, Green ED, Blakesley RW, Bouffard GG, De Jong PJ, Osoegawa K, Zhu B, Marra M, Schein J, Bosdet I, Fjell C, Jones S, Krzywinski M, Mathewson C, Siddiqui A, Wye N, McPherson J, Zhao S, Fraser CM, Shetty J, Shatsman S, Geer K, Chen Y, Abramzon S, Nierman WC, Havlak PH, Chen R, Durbin KJ, Egan A, Ren Y, Song XZ, Li B, Liu Y, Qin X, Cawley S, Cooney AJ, D'Souza LM, Martin K, Wu JQ, Gonzalez-Garay ML, Jackson AR, Kalafus KJ, McLeod MP, Milosavljevic A, Virk D, Volkov A, Wheeler DA, Zhang Z, Bailey JA, Eichler EE, Tuzun E, Birney E, Mongin E, Ureta-Vidal A, Woodwark C, Zdobnov E, Bork P, Suyama M, Torrents D, Alexandersson M, Trask BJ, Young JM, Huang H, Wang H, Xing H, Daniels S, Gietzen D, Schmidt J, Stevens K, Vitt U, Wingrove J, Camara F, Mar Alba M, Abril JF, Guigo R, Smit A, Dubchak I, Rubin EM, Couronne O, Poliakov A, Hubner N, Ganten D, Goesele C, Hummel O, Kreitler T, Lee YA, Monti J, Schulz H, Zimdahl H, Himmelbauer H, Lehrach H, Jacob HJ, Bromberg S, Gullings-Handley J, Jensen-Seaman MI, Kwitek AE, Lazar J, Pasko D, Tonellato PJ, Twigger S, Ponting CP, Duarte JM, Rice S, Goodstadt L, Beatson SA, Emes RD, Winter EE, Webber C, Brandt P, Nyakatura G, Adetobi M, Chiaromonte F, Elnitski L, Eswara P, Hardison RC, Hou M, Kolbe D, Makova K, Miller W, Nekrutenko A, Riemer C, Schwartz S, Taylor J, Yang S, Zhang Y, Lindpaintner K, Andrews TD, Caccamo M, Clamp M, Clarke L, Curwen V, Durbin R, Eyras E, Searle SM, Cooper GM, Batzoglou S, Brudno M, Sidow A, Stone EA, Payseur BA, Bourque G, Lopez-Otin C, Puente XS, Chakrabarti K, Chatterji S, Dewey C, Pachter L, Bray N, Yap VB, Caspi A, Tesler G, Pevzner PA, Haussler D, Roskin KM, Baertsch R, Clawson H, Furey TS, Hinrichs AS, Karolchik D, Kent WJ, Rosenbloom KR, Trumbower H, Weirauch M, Cooper DN, Stenson PD, Ma B, Brent M, Arumugam M, Shteynberg D, Copley RR, Taylor MS, Riethman H, Mudunuri U, Peterson J, Guyer MS, Felsenfeld A, Old S, Mockrin S, Collins F: **Genome sequence of the Brown Norway rat yields insights into mammalian evolution.** *Nature* 2004, **428:**493-521.
20. Coghlan A, Wolfe KH: **Fourfold Faster Rate of Genome Rearrangement in Nematodes Than in *Drosophila*.** *Genome Res* 2002, **12:**857-867.
21. Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, Chen N, Chinwalla A, Clarke L, Clee C, Coghlan A, Coulson A, D'Eustachio P, Fitch DH, Fulton LA, Fulton RE, Griffiths-Jones S, Harris TW, Hillier LW, Kamath R, Kuwabara PE, Mardis ER, Marra MA, Miner TL, Minx P, Mullikin JC, Plumb RW, Rogers J, Schein JE, Sohrmann M, Spieth J, Sta-

jich JE, Wei C, Willey D, Wilson RK, Durbin R, Waterston RH: **The Genome Sequence of *Caenorhabditis briggsae*: A Platform for Comparative Genomics.** *PLoS Biology* 2003, **1:**E45.
22. Zdobnov EM, von Mering C, Letunic I, Torrents D, Suyama M, Copley RR, Christophides GK, Thomasova D, Holt RA, Subramanian GM, Mueller HM, Dimopoulos G, Law JH, Wells MA, Birney E, Charlab R, Halpern AL, Kokoza E, Kraft CL, Lai Z, Lewis S, Louis C, Barillas-Mury C, Nusskern D, Rubin GM, Salzberg SL, Sutton GG, Topalis P, Wides R, Wincker P, Yandell M, Collins FH, Ribeiro J, Gelbart WM, Kafatos FC, Bork P: **Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*.** *Science* 2002, **298:**149-159.
23. Douglas S, Zauner S, Fraunholz M, Beaton M, Penny S, Deng LT, Wu X, Reith M, Cavalier-Smith T, Maier UG: **The highly reduced genome of an enslaved algal nucleus.** *Nature* 2001, **410:**1091-1096.
24. Katinka MD, Duprat S, Cornillot E, Metenier G, Thomarat F, Prensier G, Barbe V, Peyretaillade E, Brottier P, Wincker P, Delbac F, El Alaoui H, Peyret P, Saurin W, Gouy M, Weissenbach J, Vivares CP: **Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*.** *Nature* 2001, **414:**450-453.
25. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, Paulsen IT, James K, Eisen JA, Rutherford K, Salzberg SL, Craig A, Kyes S, Chan MS, Nene V, Shallom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin DMA, Fairlamb AH, Fraunholz MJ, Roos DS, Ralph SA, McFadden GI, Cummings LM, Subramanian GM, Mungall C, Venter JC, Carucci DJ, Hoffman SL, Newbold C, Davis RW, Fraser CM, Barrell B: **Genome sequence of the human malaria parasite *Plasmodium falciparum*.** *Nature* 2002, **419:**498-511.
26. Pizzi E, Frontali C: **Low-Complexity Regions in *Plasmodium falciparum* Proteins.** *Genome Res* 2001, **11:**218-229.
27. Felsenstein J: **Cases in which parsimony or compatibility methods will be positively misleading.** *Syst Zool* 1978, **27:**401-410.
28. Rokas A, Williams BL, King N, Carroll SB: **Genome-scale approaches to resolving incongruence in molecular phylogenies.** *Nature* 2003, **425:**798-804.
29. Gadagkar SR, Rosenberg MS, Kumar S: **Inferring species phylogenies from multiple genes: Concatenated seqeunce tree versus consensus gene tree.** *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 2005, **304B:**64-74.
30. Wray GA, Levinton JS, Shapiro LH: **Molecular evidence for deep Precambrian divergences among metazoan phyla.** *Science* 1996, **274:**568-573.
31. Feng DF, Cho G, Doolittle RF: **Determining divergence times with a protein clock: update and reevaluation.** *Proc Nat Acad Sci U S A* 1997, **94:**13028-13033.
32. Wang DY, Kumar S, Hedges SB: **Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi.** *Proceedings of the Royal Society of London B: Biological Sciences* 1999, **266:**163-171.
33. Nei M, Xu P, Glazko G: **Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms.** *Proc Natl Acad Sci U S A* 2001, **98:**2497-2502.
34. Hedges SB, Kumar S: **Precision of molecular time estimates.** *Trends in Genetics* 2004, **20:**242-247.
35. Peterson KJ, Lyons JB, Nowak KS, Takacs CM, Wargo MJ, McPeek MA: **Estimating metazoan divergence times with a molecular clock.** *Proc Natl Acad Sci U S A* 2004, **101:**6536-6541.
36. Douzery EJP, Snell EA, Bapteste E, Delsuc F, Philippe H: **The timing of eukaryote evolution: Does a relaxed molecular clock reconcile proteins and fossils?** *Proc Natl Acad Sci U S A* 2004, **101:**15386-15391.
37. Aris-Brosou S, Yang Z: **Bayesian models of episodic evolution support a late precambrian explosive diversification of the Metazoa.** *Molecular Biology and Evolution* 2003, **20:**1947-1954.
38. Kumar S: **Mesoproterozoic megafossil Chuaria-Tawuia association may represent parts of a multicellular plant, Vindhyan Supergroup, Central India.** *Precambrian Res* 2001, **106:**187-211.
39. Woods KN, Knoll AH, German TN: **Xanthophyte Algae from the Mesoproterozoic/Neoproterozoic Transition: Confirmation and Evolutionary Implications.** *GSA Abstracts with Programs* 1998, **30:**A232.

40. Butterfield NJ: **Bangiomorpha pubescens n. gen., n. sp.: Implications for the Evolution of Sex, Multicellularity, and the Mesoproterozoic/Neoproterozoic Radiation of Eukaryotes.** *Paleobiology* 2000, **26**:386-404.

41. Blair JE, Hedges SB: **Molecular Clocks Do Not Support the Cambrian Explosion.** *Mol Biol Evol* 2005, **22**:387-390.

42. Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B, Ferriera S, Wang G, Zheng X, White TJ, Sninsky JJ, Adams MD, Cargill M: **Inferring non-neutral evolution from human-chimp-mouse orthologous gene trios.** *Science* 2003, **302**:1960-1963.

43. Kumar S, Subramanian S: **Mutation rates in mammalian genomes.** *Proc Natl Acad Sci U S A* 2002, **99**:803-808.

44. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402.

45. Remm M, Storm CE, Sonnhammer EL: **Automatic clustering of orthologs and in-paralogs from pairwise species comparisons.** *Journal of Molecular Biology* 2001, **314**:1041-1052.

46. Tatusov RL, Galperin MY, Natale DA, Koonin EV: **The COG database: a tool for genome-scale analysis of protein functions and evolution.** *Nucleic Acids Research* 2000, **28**:33-36.

47. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X: **The sequence of the human genome.** *Science* 2001, **291**:1304-1351.

48. Ensembl Genome Browser: **http://www.ensembl.org/.** .

49. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitza-

kis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, Kent WJ, Kirby A, Kolbe DL, Korf I, Kucherlapati RS, Kulbokas EJ, Kulp D, Landers T, Leger JP, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy M, McCombie WR, McLaren S, McLay K, McPherson JD, Meldrim J, Meredith B, Mesirov JP, Miller W, Miner TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pevzner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail M, Reymond A, Roe BA, Roskin KM, Rubin EM, Rust AG, Santos R, Sapojnikov V, Schultz B, Schultz J, Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Spencer B, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Von Niederhausern AC, Wade CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang SP, Zdobnov EM, Zody MC, Lander ES: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420**:520-562.

50. Aparicio S, Chapman J, Stupka E, Putnam N, Chia J, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, Gelpke MDS, Roach J, Oh T, Ho IY, Wong M, Detter C, Verhoef F, Predki P, Tay A, Lucas S, Richardson P, Smith SF, Clark MS, Edwards YJK, Doggett N, Zharkikh A, Tavtigian SV, Pruss D, Barnstead M, Evans C, Baden H, Powell J, Glusman D, Rowen L, Hood L, Tan YH, Elgar G, Hawkins T, Venkatesh B, Rokhsar D, Brenner S: **Whole-Genome Shotgun Assembly and Analysis of the Genome of *Fugu rubripes*.** *Science* 2002, **297**:1301-1310.

51. DOE Joint Genome Institute: **http://www.jgi.doe.gov/.** .

52. Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, De Tomaso A, Davidson B, Di Gregorio A, Gelpke M, Goodstein DM, Harafuji N, Hastings KE, Ho I, Hotta K, Huang W, Kawashima T, Lemaire P, Martinez D, Meinertzhagen IA, Necula S, Nonaka M, Putnam N, Rash S, Saiga H, Satake M, Terry A, Yamada L, Wang HG, Awazu S, Azumi K, Boore J, Branno M, Chin-Bow S, DeSantis R, Doyle S, Francino P, Keys DN, Haga S, Hayashi H, Hino K, Imai KS, Inaba K, Kano S, Kobayashi K, Kobayashi M, Lee BI, Makabe KW, Manohar C, Matassi G, Medina M, Mochizuki Y, Mount S, Morishita T, Miura S, Nakayama A, Nishizaka S, Nomoto H, Ohta F, Oishi K, Rigoutsos I, Sano M, Sasaki A, Sasakura Y, Shoguchi E, Shin-i T, Spagnuolo A, Stainier D, Suzuki MM, Tassy O, Takatori N, Tokuoka M, Yagi K, Yoshizaki F, Wada S, Zhang C, Hyatt PD, Larimer F, Detter C, Doggett N, Glavina T, Hawkins T, Richardson P, Lucas S, Kohara Y, Levine M, Satoh N, Rokhsar DS: **The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins.** *Science* 2002, **298**:2157-2167.

53. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor GL, Abril JF, Agbayani A, An HJ, Andrews-Pfannkoch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV, Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC, Busam DA, Butler H, Cadieu E, Center A, Chandra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, de Pablos B, Delcher A, Deng Z, Mays AD, Dew I, Dietz SM, Dodson K, Doup LE, Downes M, Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferriera S, Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D, Houston KA, Howland TJ, Wei MH, Ibegwam C, Jalali M, Kalush F, Karpen GH, Ke Z, Kennison JA, Ketchum KA, Kimmel BE, Kodira CD, Kraft C, Kravitz S, Kulp D, Lai Z, Lasko P, Lei Y, Levitsky AA, Li J, Li Z, Liang Y, Lin X, Liu X, Mattei B, McIntosh TC, McLeod

MP, McPherson D, Merkulov G, Milshina NV, Mobarry C, Morris J, Moshrefi A, Mount SM, Moy M, Murphy B, Murphy L, Muzny DM, Nelson DL, Nelson DR, Nelson KA, Nixon K, Nusskern DR, Pacleb JM, Palazzolo M, Pittman GS, Pan S, Pollard J, Puri V, Reese MG, Reinert K, Remington K, Saunders RD, Scheeler F, Shen H, Shue BC, Siden-Kiamos I, Simpson M, Skupski MP, Smith T, Spier E, Spradling AC, Stapleton M, Strong R, Sun E, Svirskas R, Tector C, Turner R, Venter E, Wang AH, Wang X, Wang ZY, Wassarman DA, Weinstock GM, Weissenbach J, Williams SM, WoodageT, Worley KC, Wu D, Yang S, Yao QA, Ye J, Yeh RF, Zaveri JS, Zhan M, Zhang G, Zhao Q, Zheng L, Zheng XH, Zhong FN, Zhong W, Zhou X, Zhu S, Zhu X, Smith HO, Gibbs RA, Myers EW, Rubin GM, Venter JC: **The genome sequence of *Drosophila melanogaster*.** *Science* 2000, **287:**2185-2195.

54. **Berkeley *Drosophilia* Genome Project** [http://www.fruitfly.org]

55. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, Wincker P, Clark AG, Ribeiro JM, Wides R, Salzberg SL, Loftus B, Yandell M, Majoros WH, Rusch DB, Lai Z, Kraft CL, Abril JF, Anthouard V, Arensburger P, Atkinson PW, Baden H, de Berardinis V, Baldwin D, Benes V, Biedler J, Blass C, Bolanos R, Boscus D, Barnstead M, Cai S, Center A, Chaturverdi K, Christophides GK, Chrystal MA, Clamp M, Cravchik A, Curwen V, Dana A, Delcher A, Dew I, Evans CA, Flanigan M, Grundschober-Freimoser A, Friedli L, Gu Z, Guan P, Guigo R, Hillenmeyer ME, Hladun SL, Hogan JR, Hong YS, Hoover J, Jaillon O, Ke Z, Kodira C, Kokoza E, Koutsos A, Letunic I, Levitsky A, Liang Y, Lin JJ, Lobo NF, Lopez JR, Malek JA, McIntosh TC, Meister S, Miller J, Mobarry C, Mongin E, Murphy SD, O'Brochta DA, Pfannkoch C, Qi R, Regier MA, Remington K, Shao H, Sharakhova MV, Sitter CD, Shetty J, Smith TJ, Strong R, Sun J, Thomasova D, Ton LQ, Topalis P, Tu Z, Unger MF, Walenz B, Wang A, Wang J, Wang M, Wang X, Woodford KJ, Wortman JR, Wu M, Yao A, Zdobnov EM, Zhang H, Zhao Q, Zhao S, Zhu SC, Zhimulev I, Coluzzi M, della Torre A, Roth CW, Louis C, Kalush F, Mural RJ, Myers EW, Adams MD, Smith HO, Broder S, Gardner MJ, Fraser CM, Birney E, Bork P, Brey PT, Venter JC, Weissenbach J, Kafatos FC, Collins FH, Hoffman SL: **The genome sequence of the malaria mosquito *Anopheles gambiae*.** *Science* 2002, **298:**129-149.

56. The C. elegans Sequencing Consortium: **Genome sequence of the nematode *C. elegans*: a platform for investigating biology.** *Science* 1998, **282:**2012-2018.

57. Wormbase: [http://www.wormbase.org]

58. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG: **Life with 6000 genes.** *Science* 1996, **274:**546, 563-7.

59. *Saccharomyces* **Genome Database** [http://www.yeastgenome.org]

60. Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma LJ, Smirnov S, Purcell S, Rehman B, Elkins T, Engels R, Wang S, Nielsen CB, Butler J, Endrizzi M, Qui D, Ianakiev P, Bell-Pedersen D, Nelson MA, Werner-Washburne M, Selitrennikoff CP, Kinsey JA, Braun EL, Zelter A, Schulte U, Kothe GO, Jedd G, Mewes W, Staben C, Marcotte E, Greenberg D, Roy A, Foley K, Naylor J, Stange-Thomann N, Barrett R, Gnerre S, Kamal M, Kamvysselis M, Mauceli E, Bielke C, Rudd S, Frishman D, Krystofova S, Rasmussen C, Metzenberg RL, Perkins DD, Kroken S, Cogoni C, Macino G, Catcheside D, Li W, Pratt RJ, Osmani SA, DeSouza CP, Glass L, Orbach MJ, Berglund JA, Voelker R, Yarden O, Plamann M, Seiler S, Dunlap J, Radford A, Aramayo R, Natvig DO, Alex LA, Mannhaupt G, Ebbole DJ, Freitag M, Paulsen I, Sachs MS, Lander ES, Nusbaum C, Birren B: **The genome sequence of the filamentous fungus *Neurospora crassa*.** *Nature* 2003, **422:**859-868.

61. **Broad Institute** [http://www.broad.mit.edu]

62. Dietrich FS, Voegeli S, Brachat S, Lerch A, Gates K, Steiner S, Mohr C, Pohlmann R, Luedi P, Choi S, Wing RA, Flavier A, Gaffney TD, Philippsen P: **The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome.** *Science* 2004, **304:**304-307.

63. *Ashbya* **Genome Database** [http://agd.unibas.ch]

64. **European Bioinformatics Institute** [http://www.ebi.ac.uk/]

65. The *Arabidopsis* Genome Initiative: **Analysis of the genome sequence of the flowering plant *Arabidopsis* thaliana.** *Nature* 2000, **408:**796-815.

66. **The Institute for Genomic Research** [http://www.tigr.org/]

67. Matsuzaki M, Misumi O, Shin IT, Maruyama S, Takahara M, Miyagishima SY, Mori T, Nishida K, Yagisawa F, Yoshida Y, Nishimura Y,

Nakao S, Kobayashi T, Momoyama Y, Higashiyama T, Minoda A, Sano M, Nomoto H, Oishi K, Hayashi H, Ohta F, Nishizaka S, Haga S, Miura S, Morishita T, Kabeya Y, Terasawa K, Suzuki Y, Ishii Y, Asakawa S, Takano H, Ohta N, Kuroiwa H, Tanaka K, Shimizu N, Sugano S, Sato N, Nozaki H, Ogasawara N, Kohara Y, Kuroiwa T: **Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D.** *Nature* 2004, **428:**653-657.

68. *Cyanidioschyzon* **merolae Genome Project** [http://merolae.biol.s.u-tokyo.ac.jp]

69. PlasmoDB: [**http://www.plasmodb.org**]

70. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: **The CLUSTALX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.** *Nucleic Acids Research* 1997, **25:**4876-4882.

71. Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood.** *Bioinformatics* 1997, **13:**555-556.

72. Kumar S, Tamura K, Jakobsen I, Nei M: **MEGA2: molecular evolutionary genetics analysis software.** *Bioinformatics* 2001, **17:**1244-1245.

73. Strimmer K, von Haeseler A: **Quartet Puzzling: A Quartet Maximum-Likelihood Method for Reconstructing Tree Topologies.** *Mol Biol Evol* 1996, **13:**964-969.

74. Felsenstein J: **Phylip: Phylogeny Inference Package.** 3.6(a3) edition. Seattle, Department of Genome Sciences, University of Washington; 2002.

75. Kishino H, Thorne JL, Bruno WJ: **Performance of a divergence time estimation method under a probabilistic model of rate evolution.** *Molecular Biology and Evolution* 2001, **18:**352-361.

76. Jacobs LL, Downs WR: **The evolution of murine rodents in Asia.** In *Rodent and Lagomorph Families of Asian Origins and Diversification Volume 8*. Edited by: Tomida Y, Li CK and Setoguchi T. Tokyo, National Science Museum Monographs; 1994:149-156.

77. Benton MJ: **The Fossil Record 2.** New York, Chapman & Hall; 1993.

78. Donoghue PCJ, Smith MP, Sansom IJ: **The origin and early evolution of chordates: molecular clocks and the fossil record.** In *Telling the Evolutionary Time: Molecular Clocks and the Fossil Record* Edited by: Donoghue PCJ and Smith MP. New York, CRC Press; 2004:190-223.

79. Shu DG, Chen L, Han J, Zhang XL: **An Early Cambrian tunicate from China.** *Nature* 2001, **411:**472-473.

80. Gaunt MW, Miles MA: **An insect molecular clock dates the origin of the insects and accords with palaeontological and biogeographic landmarks.** *Molecular Biology and Evolution* 2002, **19:**748-761.